Recursive Equations Based Models of Queueing Systems*

Nikolai K. Krivulin†

Abstract

An overview of the recursive equations based models and their applications in simulation based analysis and optimization of queueing systems is given. These models provide a variety of systems with a convenient and unified representation in terms of recursions for arrival and departure times of customers, which involves only the operations of maximum, minimum, and addition.

1 Introduction

As a representation of dynamics of queueing systems, recursive equations have been introduced by Lindley 1952 in his investigation of the $G/G/1$ queue. The representation has proved to be useful in both analytical study and simulation of queues, and was extended to cover a variety of queueing systems including open and closed tandem single-server queues with both infinite and finite buffers, the $G/G/m$ system, and queueing networks.

The recursive equations were originally expressed in terms of the waiting times of customers ([1, 2]). Equations following this classical approach remain traditional in the queueing theory, one can find them in many of the recent works devoted mainly to theoretical aspects of the investigation of queueing systems (see, e.g., [3]).

In the last few years, another representation based on recursions for the arrival and departure times of customers has gained acceptance in works dealing with the simulation study of queueing systems and its related fields including performance evaluation and sensitivity analysis. The items of our list of references, other than those cited above, can serve as an illustration. Although these equations may be readily derived from those of the classical type, they offer a more convenient and unified way of representing dynamics of queueing systems as well as their performance measures.

*Proc. 1994 SCS Europ. Simulation Symp., Istanbul, Turkey, Oct. 9-12, 1994 / Ed. by A. R. Kaylan, A. Lehmann, T. I. Oren. P. 252–256.

†Faculty of Mathematics and Mechanics, St. Petersburg State University, Bibliotechnaya Sq.2, Petrodvorets, St. Petersburg, 198904 Russia

The purpose of this paper is to give a brief overview of the recursive equations and their applications in simulation based analysis and optimization of queueing systems. The subsequent sections present the equations expressed in terms of the arrival and departure times, which describe the systems most commonly encountered in studies of queues, and discuss the representation of performance measures, associated with these queueing system models. Applications of the models to the development of simulation algorithms as well as to the analysis of system performance measures and estimation of their sensitivity are also outlined. Finally, limitations on the use of the models are briefly discussed.

2 Models of Queueing Systems

Most of the models appearing in this section actually present single-server systems which can have both finite and infinite buffers, and operate according to the first-come, first-served (FCFS) queueing discipline. Also included are the equations representing the $G/G/m$ system, and a rather general model of a queueing network with a deterministic routing mechanism.

2.1 The $G/G/1$ Queue

We start with this model which provides the basis for representing more complicated queueing systems. The $G/G/1$ system consists of a server and a buffer with infinite capacity (Fig.1). Once a customer arrives into the system, he occupies the server provided that it is free. If the customer finds the server busy, he is placed into the buffer, and starts waiting to be served. The queue discipline in the system is presumed to be FCFS.

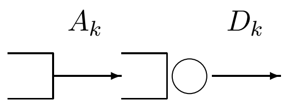

Figure 1: The $G/G/1$ queue.

For the $G/G/1$ queue, we denote the interarrival time between the kth customer and his predecessor by α_k, and the service time of the kth customer by τ_k. Furthermore, let A_k be the kth arrival epoch to the queue, and D_k be the kth departure epoch from the queue, $k = 1, 2, \ldots$. Provided that the system starts operating at time zero, it is convenient to set $A_0 \equiv 0$ and $D_0 \equiv 0$. One may now describe the dynamics of the $G/G/1$ queue as

$$\begin{aligned} A_k &= A_{k-1} + \alpha_k \\ D_k &= (A_k \vee D_{k-1}) + \tau_k, \end{aligned}$$

where $\vee$ stands for the maximum operator, $k = 1, 2, \ldots$.

2.2 Tandem Systems of Single-Server Queues

Consider a system of N single-server queues with infinite buffers, operating in tandem as shown in Fig.2.

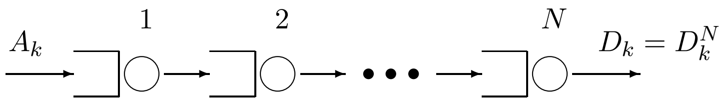

Figure 2: Single-server queues in tandem.

Each customer that arrives into this system has to pass through all the queues so as to occupy consecutively every servers from 1 to N, and then leave the system. We suppose that upon his service completion at a queue, the customer arrives into the next queue immediately.

To set up the recursive equations representing the system in a convenient way, let us introduce the symbols D_k^n and τ_k^n respectively for the departure and service times of the kth customer at queue n. However, we maintain the symbols A_k and $D_k = D_k^N$ to denote the kth arrival and departure epochs for the whole system. With these notations, the equations are written as (Shanthikumar and Yao 1989a; and Chen and Chen 1990)

$$\begin{aligned} D_k^1 &= (A_k \vee D_{k-1}^1) + \tau_k^1 \\ D_k^n &= (D_k^{n-1} \vee D_{k-1}^n) + \tau_k^n, \quad n = 2, \ldots, N. \end{aligned}$$

2.2.1 Closed Tandem Systems

Suppose that in the above tandem system all the customers after their service completion at the Nth server return to the 1st queue for the next cycle of service (see Fig.3).

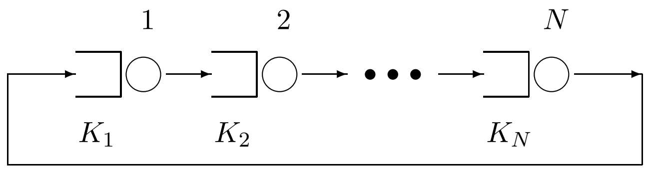

Figure 3: A closed tandem system of single-server queues.

Furthermore, we assume that at the initial time, there are K_n, $0 \leq K_n < \infty$, customers in the buffer of server n. Assuming $D_k^n = -\infty$ for all

$k < 0$ and $n = 1, \dots N$, one can represent the closed system in the form (see, e.g., [4, 5])

$$\begin{aligned} D_k^1 &= (D_{k-K_1}^N \vee D_{k-1}^1) + \tau_k^1 \\ D_k^n &= (D_{k-K_n}^{n-1} \vee D_{k-1}^n) + \tau_k^n, \quad n = 2, \dots, N. \end{aligned}$$

2.2.2 Tandem Queues with Finite Buffers

We now turn to the discussion of the system of queues which provide only a limited number of places in their buffers for customers waiting for service. In such a system, if the buffer at a server has finite capacity, the preceding server may be blocked according to one of the blocking rules. In this paper we shall restrict ourselves to *manufacturing* blocking and *communication* blocking which are more frequent in practice.

Consider a system of N queues, depicted in Fig.4. We denote the capacity of the buffer at server n by B_n, $0 \leq B_n \leq \infty$, $n = 2, 3, \dots, N$. As the input buffer of the system, the buffer of the 1st server is assumed to be infinite.

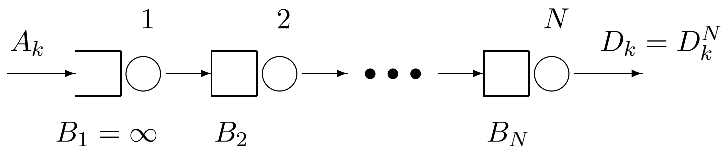

Figure 4: Tandem single-server queues with finite buffers.

Let us first suppose that the system operates according to the manufacturing blocking rule. Under this type of blocking, if a customer upon completion of his service at server n sees the buffer of the $(n+1)$st server full, he cannot unoccupy the nth server until the next server provides a free space in its buffer. Since buffers become free as customers are called forward for service, the nth server is unoccupied as soon as the $(n+1)$st server completes its current service to initiate the service of the next customer. It is not difficult to understand that the departure of the kth customer from server n occurs not earlier than that of the $(k - B_{n+1} - 1)$st customer from server $n+1$. Taking into account this condition, one may represent the equations as ([6, 7])

$$\begin{aligned} D_k^1 &= ((A_k \vee D_{k-1}^1) + \tau_k^1) \vee D_{k-B_2-1}^2 \\ D_k^n &= ((D_k^{n-1} \vee D_{k-1}^n) + \tau_k^n) \vee D_{k-B_{n+1}-1}^{n+1}, \quad n = 2, \dots, N-1 \\ D_k^N &= (D_k^{N-1} \vee D_{k-1}^N) + \tau_k^N. \end{aligned}$$

The communication blocking rule requires from a server not to initiate the service of a customer if the buffer of the next server is full. In this case,

the server remains unavailable until the current service at the next server is completed. For the system with communication blocking, we have (Chen and Chen 1990)

$$
\begin{aligned}
D_k^1 &= (A_k \vee D_{k-1}^1 \vee D_{k-B_2-1}^2) + \tau_k^1 \\
D_k^n &= (D_k^{n-1} \vee D_{k-1}^n \vee D_{k-B_{n+1}-1}^{n+1}) + \tau_k^n, \quad n = 2, \dots, N-1 \\
D_k^N &= (D_k^{N-1} \vee D_{k-1}^N) + \tau_k^N.
\end{aligned}
$$

2.3 $G/G/m$ Queues

Equations representing the $G/G/m$ queue (Fig.5) as recursions for the waiting times of customers have been first introduced by Kiefer and Wolfowitz 1955 [2]. These recursive equations were expressed in general terms rather than in an explicit form similar to those presented above.

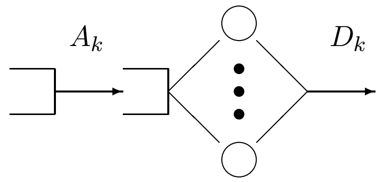

Figure 5: The $G/G/m$ queue.

To represent the equations for the $G/G/m$ queue, $1 \leq m < \infty$, in terms of the arrival and departure times of customers, let us further insert the symbol C_k for the service completion time of the customer which is the kth to arrive into the system. Note that in multi-server queues the kth departure time and the completion time of the kth customer may not coincide as contrasted to the $G/G/1$ queue which does not recognize them.

Now we may describe the dynamics of the system through the equations proposed in [8] (a similar representation in terms of waiting times can be found in [9])

$$
\begin{aligned}
A_k &= A_{k-1} + \alpha_k \\
C_k &= (A_k \vee D_{k-m}) + \tau_k \\
D_k &= \bigwedge_{1 \leq j_1 < \dots < j_k \leq k+m-2} (C_{j_1} \vee \cdots \vee C_{j_k}) \wedge C_{k+m-1},
\end{aligned}
$$

where $\wedge$ signifies the minimum operator. Note that with $m = 1$ the above set of equations is reduced to that of the $G/G/1$ queue.

2.4 Networks with Deterministic Routing

We complete this section by presenting a rather general model of a closed queueing network with deterministic routing described in [10, 11] (see also a

similar model in [12]). Let us first consider a network consisting of N single-server nodes. In each node there are a server and an infinite buffer in which customers are placed at their arrival so as to wait for service if it cannot be initiated immediately. After his service completion at one node, each customer goes to another node chosen according to the routing procedure defined as follows. For the network, we assume that a matrix

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1k} & \cdots \\ s_{21} & s_{22} & \cdots & s_{2k} & \cdots \\ \vdots & \vdots & & \vdots & \\ s_{N1} & s_{N2} & \cdots & s_{Nk} & \cdots \end{pmatrix}$$

is given, s_{nk} determines the next node to be visited by the customer who is the kth to depart from node n, $s_{nk} \in \{1, \ldots, N\}$; $n = 1, \ldots, N$; $k = 1, 2, \ldots$. It is also assumed that at the initial time, all servers are free, and there are K_n, $0 \leq K_n \leq \infty$, customers in the buffer at node n.

For node n, we denote the kth arrival and departure epochs respectively by A_k^n and D_k^n, and the service time of the customer who is the kth to arrive by τ_k^n. Furthermore, let us introduce the set

$$\boldsymbol{D}_n = \{D_i^k | s_{ik} = n; i = 1, \ldots, N; k = 1, 2, \ldots\},$$

which is constituted by the departure times of the customers who have to go to node n. Finally, we may represent the network by means of the equations

$$\begin{aligned} D_k^n &= (A_k^n \vee D_{k-1}^n) + \tau_k^n \\ A_k^n &= \begin{cases} 0, & \text{if } k \leq K_n \\ \mathcal{A}_{k-K_n}^n, & \text{otherwise,} \end{cases} \end{aligned}$$

where $\mathcal{A}_k^n$ is the arrival time of the customer which is the kth to arrive into node n after his service at any node of the network. In other words, the symbol $\mathcal{A}_k^n$ differs from A_k^n in that it relates only to the customers really arriving into node n, and does not to those occurring in this node at the initial time. It is defined as

$$\mathcal{A}_k^n = \bigwedge_{\{D_1, \ldots, D_k\} \subset \boldsymbol{D}_n} (D_1 \vee \cdots \vee D_k),$$

where minimum is taken over all k-subsets of the set $\boldsymbol{D}_n$.

It is easy to understand how tandem queueing systems with infinite buffers may be represented as networks like that just described. Moreover, changing the first one from the above equations, one can readily extend the model to cover networks with nodes which may have many servers. These servers may operate both in tandem and in parallel, and even form a network themselves.

3 Performance Measures

We are now in a position to show how the performance measures which one normally chooses in the analysis of queueing systems may be represented on the basis of the models described above. We start with presenting sample performance measures associated with the systems under consideration, and then briefly discuss the evaluation of system performance measures in the general case.

3.1 Networks with Single-Server Nodes

Suppose that we observe the network until the Kth service completion at node n, $K = 1, 2, \ldots;\ 1 \leq n \leq N$. As sample performance measures for node n in the observation period, one can consider the following average quantities ([7, 10, 11]):

$$\begin{aligned}
S_K^n &= \sum_{k=1}^{K} (D_k^n - A_k^n)/K, \quad \text{the total time of one customer;} \\
W_K^n &= \sum_{k=1}^{K} (D_k^n - A_k^n - \tau_k^n)/K; \quad \text{the waiting time of one customer;} \\
T_K^n &= K/D_K^n, \quad \text{the throughput rate of the node;} \\
U_K^n &= \sum_{k=1}^{K} \tau_k^n / D_K^n, \quad \text{the utilization of the server;} \\
J_K^n &= \sum_{k=1}^{K} (D_k^n - A_k^n)/D_K^n, \quad \text{the number of customers at the node;} \\
Q_K^n &= \sum_{k=1}^{K} (D_k^n - A_k^n - \tau_k^n)/D_K^n, \quad \text{the queue length at the node.}
\end{aligned}$$

Note that, assuming the service times τ_k^n to be given, one can express these measures in closed form only in terms of these times, involving arithmetic operations, and the operations of maximum and minimum ([11]).

3.2 Tandem Systems of Queues

Since tandem systems can be considered as networks with deterministic routing, the above sample performance measures are also suited to the tandem systems. In addition to these measures which are actually server related performance criteria, for a tandem system with N servers we may define

customer related performance measures [7]

$$
\begin{aligned}
S_K &= \sum_{k=1}^{K} (D_k - A_k)/K, \quad \text{the average system time of one customer;} \\
W_K &= \sum_{k=1}^{K} \left(D_k - A_k - \sum_{n=1}^{N} \tau_k^n \right) /K, \quad \text{the average waiting time of one customer.}
\end{aligned}
$$

Finally, there are sample performance measures inherent only in the systems with finite buffers. As an example, the average idle time of a server, say server n, can be considered. This measure is written in the same form for both the manufacturing and communication blocking rules as

$$
I_K^n = \sum_{k=1}^{K} \left(D_k^n - (D_k^{n-1} \vee D_{k-1}^n) - \tau_k^n \right) /K.
$$

3.3 Multi-Server Queues

Sample performance measures in multi-server queueing systems can be represented through formulas which are closely similar to those applied in queueing networks. For instance, the average throughput may be defined exactly as we have defined T_K^n. To represent properly the remaining measures, one however has to take into account the distinction between the kth completion and the kth departure times, involved in the $G/G/m$ queue. With this distinction, replacing the symbols D_k by C_k is required in the above formulas so as to provide appropriate expressions for the sample performance measures of multi-server queues.

3.4 Evaluation of System Performance

We suppose now that the service times τ_k (and the interarrival times α_k if they are given) are defined as random variables $\tau_k = \tau_k(\theta, \omega)$, where $\theta \in \Theta$ is a set of decision parameters, and ω is a random vector. In this case, as it results from the above representations of the queueing systems and their performance, the arrival epochs A_k and the departure epochs D_k, together with the sample performance measures also present random variables. Let $F_K = F_K(\theta, \omega)$, be a sample performance measure of the system. As is customary, we define the system performance measure associated with F_K by the expected value

$$
\boldsymbol{F}_K(\theta) = E_\omega[F_K(\theta, \omega)].
$$

Based on a finite observation period, $\boldsymbol{F}_K$ is generally referred to as *finite-horizon* performance measure. Another criterion, *a steady-state* performance

measure, intended to describe a long time behaviour of a system is defined as

$$\boldsymbol{F}(\theta) = \lim_{K\to\infty} E_\omega[F_K(\theta,\omega)].$$

Although we may express sample performance measures in closed form, in the case of general random variables determining the service times of customers, it is often very difficult or even impossible to obtain analytically the criteria $\boldsymbol{F}_K$ and especially $\boldsymbol{F}$. In this situation, one generally applies a simulation technique which allows of obtaining values of $F_K(\theta,\omega)$, and then estimates the system performance by using the Monte Carlo approach. Note however, that information concerning the explicit form of the sample performance measures normally proves to be very useful to the simulation study and optimization of queueing systems.

4 Application of the Models

In this section we briefly outline a selection of the application areas of the recursive representation in simulation based analysis and optimization of queueing systems. The section concludes with remarks concerning limitations on the use of the models in representing queueing systems.

4.1 Design of Simulation Algorithms

Since recursive equations determine a global structure of changes in queueing systems consecutively in a very natural way, they provide the basis for the development of very efficient simulation procedures (see, e.g., [7, 4, 13]). Although the simulation technique based on recursive representations of queueing systems may rank below the traditional event-scheduling approach in its versatility, the algorithms applying this technique are normally superior to others in reducing time and memory costs. Moreover, these algorithms are usually best suited to the implementation on parallel and vector processors. As an illustration, one can consider parallel simulation algorithms in [4, 13].

4.2 Variance Reduction in Simulation

Closely related to the queueing system simulation procedures are variance reduction techniques which are intended to improve the accuracy of simulation output ([14]). In order for a variance reduction method to be successfully employed in estimating a system performance $\boldsymbol{F}_K$, certain conditions normally have to be imposed on its associated sample performance measure F_K. Specifically, *the antithetic variates* method and *the common random numbers* method require that F_K as a function of the random argument ω be monotone (see, e.g., [14]). Examples of establishing such monotonicity properties from the recursive representation of queueing systems can be found in [15].

4.3 Investigation of System Performance Measures

Another area of applications of the models includes the investigation of properties inherent in performance measures of queueing systems, such as monotonicity and convexity with respect to system parameters θ. It is normally not difficult to examine these properties for the systems described by equations involving only the operations of maximum and addition (e.g., tandem queues with both infinite and finite buffers). One can find an extended discussion of this subject in [6, 12, 16].

4.4 Sensitivity Analysis and Estimation

Since there are generally no explicit representations as functions of system parameters θ available for the performance measure, one may evaluate its sensitivity (or its gradient, when the parameters are continuous) by no way other than through the use of estimates obtained from simulation experiments. Very efficient procedures of obtaining gradient estimates may be designed using new technique called *infinitesimal perturbation analysis* (IPA) (see, e.g., [17]). The IPA algorithms which are actually based on the recursive representations of queueing systems, can serve as an important line of the application of the models under discussion ([18, 11]). Finally, these models provided a useful framework for examining unbiasedness and consistency of IPA estimates in [19, 16, 10, 11].

4.5 Limitations on the Use of the Models

One can see that the general model of the network, as it has been presented above, treats of queueing systems from the viewpoint of service facilities rather than of particular customers. Specifically, for each node n only the arrival and departure instants are essential, whereas it makes no difference which of the customers proves to arrive or to depart. Moreover, both times A_k^n and D_k^n do not need to be associated with a single customer, as it normally happens in nodes with many servers operating in parallel. As a consequence, the models do not allow of representing systems with many classes of customers through recursive equations in closed form. Finally, since nodes do not distinguish among customers in some sense, the order in which customers are selected from a queue for service is of no concern, and therefore, these models are incapable of identifying distinct queue disciplines.

References

[1] D. V. Lindley, "The theory of queues with a single server," *Math. Proc. Cambridge Philos. Soc.* **48** no. 2, (1952) 277–289.

[2] J. Kiefer and J. Wolfowitz, "On the theory of queues with many servers," *Trans. Amer. Math. Soc.* **78** no. 1, (1955) 1–18.

[3] V. V. Kalashnikov, *Mathematical Methods in Queueing Theory.* Kluwer Academic Publishers, Dordrecht, 1994.

[4] A. G. Greenberg, B. D. Lubachevsky, and I. Mitrani, "Algorithms for unboundedly parallel simulations," *ACM Trans. Comput. Syst.* **9** no. 3, (1991) 201–221.

[5] N. K. Krivulin, "Using max-algebra linear models in the representation of queueing systems," in *Proc. 5th SIAM Conf. on Applied Linear Algebra*, J. G. Lewis, ed., pp. 155–160. SIAM, Philadelphia, 1994. `arXiv:1210.6019 [math.OC]`.

[6] J. G. Shanthikumar and D. D. Yao, "Second-order stochastic properties in queueing systems," *Proc. IEEE* **77** no. 1, (1989) 162–170.

[7] L. Chen and C.-L. Chen, "A fast simulation approach for tandem queueing systems," in *1990 Winter Simulation Conference Proceedings*, O. Balci, R. P. Sadowski, and R. E. Nance, eds., pp. 539–546. IEEE, 1990.

[8] N. K. Krivulin, "A recursive equations based representation for the g/g/m queue," *Appl. Math. Lett.* **7** no. 3, (1994) 73–77, `arXiv:1210.6012 [math.OC]`.

[9] G. Tsitsiashvili, "Piecewise-linear markov chains and analysis of their stability," *Theory Probab. Appl.* **20** no. 2, (1976) 337–350.

[10] N. K. Krivulin, "Optimization of complex systems for simulation modeling," *Vestnik Leningrad Univ. Math.* **23** no. 2, (1990) 64–67.

[11] N. Krivulin, "Unbiased estimates for gradients of stochastic network performance measures," *Acta Appl. Math.* **33** (1993) 21–43, `arXiv:1210.5418 [math.OC]`.

[12] J. G. Shanthikumar and D. D. Yao, "Stochastic monotonicity in general queueing networks," *J. Appl. Prob.* **26** no. 2, (1989) 413–417.

[13] S. M. Ermakov and N. K. Krivulin, "Efficient algorithms for tandem queueing system simulation," *Appl. Math. Lett.* **7** no. 6, (1994) 39–43, `arXiv:1210.6378 [math.NA]`.

[14] R. C. H. Cheng, "Variance reduction methods," in *1986 Winter Simulation Conference Proceedings*, pp. 539–546. 1986.

[15] R. Y. Rubinstein, G. Samorodnitsky, and M. Shaked, "Antithetic variates, multivariate dependence and simulation of stochastic systems," *Management Sci.* **31** no. 1, (1985) 66–77.

[16] J.-Q. Hu, "Convexity of sample path performance and strong consistency of infinitesimal perturbation analysis estimates," *IEEE Trans. Automat. Control* **37** no. 2, (1992) 258–262.

[17] Y.-C. Ho, "Performance evaluation and perturbation analysis of discrete event dynamic systems," *IEEE Trans. Automat. Control* **32** no. 7, (1987) 463–572.

[18] C. G. Cassandras and S. G. Strickland, "Perturbation analytic methodologies for design and optimization of communication networks," *IEEE J. Sel. Areas Commun.* **6** no. 1, (1988) 158–171.

[19] C. Cassandras, W. Gong, and J. Lee, "Robustness of perturbation analysis estimators for queueing systems with unknown distributions," *J. Optim. Theory Appl.* **70** (1991) 491–519.